Learning the grounding of expressions for spatial relations between objects

Mota, Tiago ; Sridharan, Mohan

In the Workshop on *Perception, Inference and Learning for Joint Semantic, Geometric and Physical Understanding* at ICRA, Brisbane, Australia, May 21, 2018.

Learning the Grounding of Expressions for Spatial Relations between Objects

Tiago Mota
Electrical and Computer Engineering
The University of Auckland, NZ
tmot987@aucklanduni.ac.nz

Mohan Sridharan
School of Computer Science
University of Birmingham, UK
m.sridharan@bham.ac.uk

***Abstract*— Robots interacting with humans often have to recognize, reason about and describe the spatial relations between objects. Prepositions are often used to describe such spatial relations, but it is difficult to equip a robot with comprehensive knowledge of these prepositions. This paper describes an architecture for incrementally learning and revising the grounding of spatial relations between objects. Answer Set Prolog, a declarative language, is used to represent and reason with incomplete knowledge that includes prepositional relations between objects in a scene. A generic grounding of prepositions for spatial relations, human input (when available), and non-monotonic logical inference, are used to infer spatial relations in 3D point clouds of given scenes, incrementally acquiring and revising a specialized metric grounding of the prepositions, and learning the relative confidence associated with each grounding. The architecture is evaluated on a benchmark dataset of tabletop images and on complex, simulated scenes of furniture.**

I. INTRODUCTION

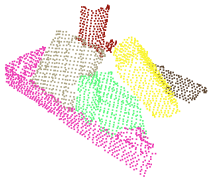
Robots[1] deployed to assist humans in complex domains such as offices and warehouses need to perform a variety of tasks in the presence of unreliable sensing and actuation, and incomplete knowledge of domain objects and relations between them. These problems are partially offset by the robot's ability to sense and interact with the environment and humans, using the corresponding observations to revise the existing knowledge. Since humans may not have the time or expertise to provide comprehensive feedback, the robot can learn more effectively by referring to objects or events of interest in terms of other known objects. For instance, in Figure 1a, asking "what is behind the cereal box?" will quickly direct the human's attention to the box of crisps. In this paper, we focus on reasoning with spatial relations between objects, and incrementally learning the "grounding" (i.e., meaning in the physical world) of words used to describe these relations.

Spatial relations are often described using prepositions, i.e., words such as *above*, *below*, *behind*, and *in*. To reason with these prepositions, the robot needs both a vocabulary and a *grounding* of these words, e.g., a mapping of these words to 3D regions, or distances from reference points or objects of interest. However, this grounding has to be revised over time in dynamic domains to account for errors, changes in viewpoint etc. Any errors in the grounding may add incorrect information to the agent's knowledge, resulting in decisions and plans that are incorrect or sub-optimal. The architecture described in this paper seeks to address these challenges and has the following key characteristics:

- A declarative language is used to represent incomplete domain knowledge, including spatial relations between objects computed using a generic (initial) grounding of prepositions in the 3D regions around objects.
- Non-monotonic logical inference with the existing knowledge, and human input (when available), are used to infer spatial relations in point clouds of new scenes, incrementally learning a specialized, histogram-based grounding of prepositions.
- Human input (when available) is used to incrementally compute the relative accuracy of spatial relations inferred by the generic and specialized groundings, using the more reliable grounding in subsequent scenes.

In this paper, we use Answer Set Prolog (ASP) as the declarative language. We consider point clouds of objects in a scene, e.g., Figure 1b, as the input and include prepositions for seven position-based and three distance-based relations. We do not explicitly represent the uncertainty in processing visual input; any conclusion drawn with high probability is elevated to a logic statement with complete certainty. The key capabilities are to enable robots to (a) start inferring spatial relations using a generic, manually-encoded grounding; (b) incrementally learn a specialized grounding of spatial relations from a small number of examples; and (c) determine the relative trust in each grounding and use the more reliable grounding for subsequent inference. We evaluate these capabilities on a benchmark dataset of tabletop objects and complex, simulated scenes of furniture.

(a) Example image of scene.



(b) Point clouds of scene.

Fig. 1: (a) Illustrative image of scene with objects; and (b) segmented version with 3D point clouds of objects in different colors.

[1]Terms "robot" and "agent" are used interchangeably.

II. RELATED WORK

Approaches found in the related literature for grounding and interpreting the spatial relations between objects are broadly based on the use of manually encoded rules, or the use of training or learning algorithms. When rules are manually encoded, the construction of a spatial vocabulary is often based on *Qualitative Spatial Representations* (QSR) such as [1], [2], [3]. These approaches may not provide accurate estimates of the spatial relations as they often approximate objects as points or establish rigid boundaries between spatial relations. Moreover, the spatial relations are encoded in advance, whereas the interpretations of these relations are likely to change over time in robotics domains. Approaches that seek to train or learn the spatial relations or their grounding do so based on *Metric Spatial Representations* (MSR), i.e., set of measures such as angles and distances between objects. Algorithms based on MSR have been used in different applications in recent years. For example, an approach based on MSR has been developed to predict the success of a robot's action in a previously unseen scenario [4], while another approach enabled an agent to learn relations between objects and generalize them to new objects [5]. Other work has focused on developing a system capable of choosing appropriate prepositions to describe an image [6]. In the context of human-robot interaction, a system has been developed for executing a set of actions on objects and answering queries about spatial positions [7], QSR and MSR have been compared for scene understanding [8], and MSR and a kd-tree have been used to dynamically infer spatial relations between objects [9]. However, most of these approaches learn the representation of spatial relations offline or in a separate training phase. In contrast, we propose an approach that initially applies a hand-designed generic grounding, and incrementally and interactively learns a histogram-based specialized grounding from new experiences and feedback.

In recent years, there has been a lot of work on inferring spatial relationships from images and natural language descriptions for tasks such as navigation and manipulation, using neural network (or deep) architectures [10], [11], [12]. These methods require a large number of training examples, learn the grounding offline, and are computationally expensive. Our architecture, on the other hand, combines the complementary strengths of non-monotonic logical inference, QSR, MSR and interactive learning to reliably and efficiently ground the spatial relations from a small number of images and through effective use of human input.

III. ARCHITECTURE

Figure 2 shows the key components of the architecture. We consider seven position-based prepositions (*in, above, below, front, behind, right, left*) and three distance-based prepositions (*touching, not-touching, far*), which are used to encode spatial relations between specific scene objects as logic statements in Answer Set Prolog (ASP), a declarative programming paradigm. The QSR module provides the initial (manually-encoded), generic grounding of spatial relations,

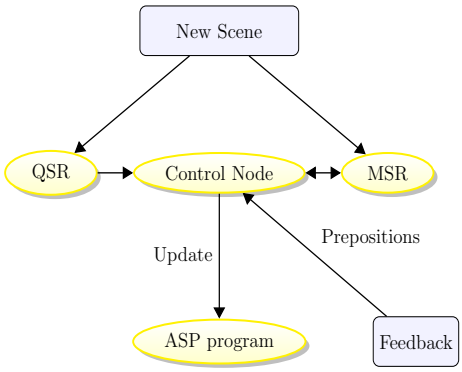
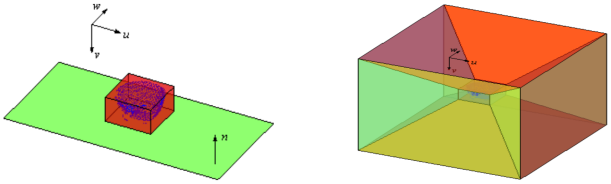


Fig. 2: Proposed architecture.



(a) Bounding Box. (b) Six Pyramids.

Fig. 3: (a) Bounding box for point cloud of a particular object; and (b) Pyramids delimiting space around the bounding box.

which is used to extract spatial relations between pairs of 3D point clouds in an input scene (the "new observation"). Human input (when available) is also used to label relations between pairs of objects. The QSR-based output and/or human input are used by the MSR module to incrementally ground the prepositions as histograms. The control node also assumes human feedback to be accurate and incrementally computes the relative confidence in the QSR and MSR groundings. The more reliable grounding is used to extract logic statements (to be added to ASP program) from subsequent images. These components are described below, but components that are not the focus of this work are not described, e.g., we sub-sample the 3D point cloud of a scene and use the Euclidean cluster extraction segmentation algorithm [13][2] to segment the point cloud.

A. Qualitative Spatial Representation

Our QSR model is similar to that proposed by [2]. For any given 3D point cloud, a bounding box containing it (i.e., around the reference object) is created—see Figure 3a; any other object with most of its point cloud located inside this bounding box is considered to be *in* the reference object. Then, the space around the reference object is divided into pyramids representing *left, right, front, behind, above,* and *below*—see Figure 3b. This definition of *in* leads to errors, especially in domains with non-convex objects. For example, if a large table is the reference object, a book that is *under* the table may be classified (incorrectly) as being *in* the table because the bounding box of the table envelopes most of the point cloud corresponding to the book.

For ease of representation, our approach differs from [2] in the definition of the distance-related prepositions: *touching,*

[2]Available at `www.pointclouds.org` for download.

not-touching, far. For a pair of point cloud clusters, we filter the 10% closest distances between pairs of points drawn from the two point clouds, and used the following criteria to determine if the two objects were touching, not touching, or distinctly separated (i.e., far) from each other:

$$\begin{aligned} &touching \;\Rightarrow\; distance(10\%) \;\leq\; 0.01 \\ &not\text{-}touching \;\Rightarrow\; 0.01 < distance(10\%) \;< 1.0 \\ &far \;\Rightarrow\; distance(10\%) \;\geq\; 1.0 \end{aligned} \tag{1}$$

where distances are measured in meters. Recall that the generic, manually-encoded grounding based on this QSR model is used by the robot to identify spatial relations between objects. This QSR model does not change over time and is based on the reasonable assumption that the robot has some initial idea of its camera's pose with respect to the scene. Next, we describe a specialized grounding that can be learned using the QSR model and/or human input.

B. Metric Spatial Representation

Unlike the QSR-based grounding, the MSR-based grounding is acquired incrementally from the input images using QSR-based labels and human feedback in the form of textual labels describing spatial relations between pairs of scene objects. Assume temporarily that the MSR module receives a pair of point clouds (corresponding to two objects) and the prepositions of the spatial relations between the objects. Each preposition is grounded using histograms, henceforth also referred to as "visual words" that are created by considering the point cloud data in a spherical coordinate system. Each point is represented by its distance to a reference point and two angles (i) $\theta \in [0°, 180°]$; and (ii) $\varphi \in [-180°, 180°]$. The coordinate frame for grounding can be based on the robot's coordinate frame, its camera, and/or reference objects—information in one frame can easily be mapped to another. Non-monotonic logical reasoning and incremental learning support recovery from errors made due to noise in sensor input processing.

We ground each of the seven position-based prepositions as 2D histograms of angles θ and φ, whereas each of the three distance-based prepositions are ground using 1D histograms of the 10% closest distances between points in pairs of objects. Figure 4 shows an example of a 2D position-based histogram. All histograms are normalized to ensure that large objects containing many points do not have an undue influence on the grounding of relations.

Once a MSR-based grounding has been learned for one or more spatial relations, it can be updated and used on new scenes. For any given pair of point cloud clusters in any new scene, the corresponding 2D and 1D histograms are constructed and compared with the learned visual words. The learned visual word most similar to the visual words extracted from new scene is used to specify the distance-based and position-based spatial relations between the two objects, e.g., "object X is below object Y and not touching it"; such statements are translated automatically to statements in the ASP program. Since axioms in the ASP program are applied recursively, each point cloud cluster needs to be considered only once. The similarity between visual words is computed using the *intersection* measure for 1D (distance) histograms. For the 2D (position) histograms, we use the χ^2 measure, e.g., for any two histograms H and G:

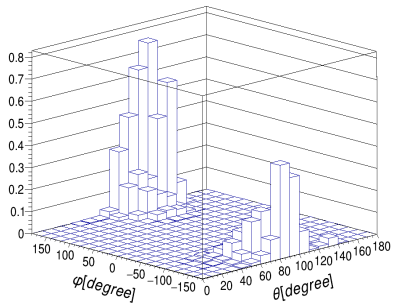


Fig. 4: Example of 2D position histogram grounding "left".

$$D_{\chi^2}(H, G) = \sum_i \frac{|h_i - g_i|^2}{2(h_i + g_i)} \tag{2}$$

where h_i and g_i are bins in H and G respectively; larger values denote greater similarity. We use this measure for 2D histograms because the boundaries between the position-based relations are more difficult to define than those between distance-based relations. Once the spatial relations between a pair of point cloud clusters have been determined in a new scene, this information updates the visual words using a standard normalized histogram merging approach, i.e., *the MSR-based grounding is updated continuously*.

C. Domain Representation in ASP

To represent and reason with incomplete knowledge, we use Answer Set Prolog (ASP), a declarative language that can represent recursive definitions, defaults, causal relations, special forms of self-reference, and language constructs that occur frequently in non-mathematical domains, and are difficult to express in classical logic formalisms. ASP is based on the stable model semantics of logic programs [14].

An ASP *program* (Π) has a *sorted signature* Σ and axioms. The signature includes sorts such as $object$, $location$, $color$, $shape$, and $step$ (to reason over time), *statics*, i.e., domain attributes that do not change over time, and *fluents*, i.e., domain attributes whose truth values can be changed. In our case, the spatial relations are fluents such as:

$$\begin{aligned} in(object, object), \quad above(object, object), \\ touching(object, object), \quad left(object, object). \end{aligned} \tag{3}$$

which are described in terms of their arguments' sorts. In addition, predicate `holds(fluent, step)` implies that a particular fluent holds true at a particular timestep.

Axioms encode some rules and relations that build on the spatial relations whose grounding is learned:

$$\begin{aligned} holds(above(A, B), I) \leftarrow & holds(below(B, A), I). \\ holds(under(A, B), I) \leftarrow & holds(touch(A, B), I), \\ & holds(below(A, B), I). \end{aligned} \tag{4}$$

where the second axiom says that any object A that is below object B and touching it is considered to be $under$ it. When action effects are to be modeled, the signature and axioms include *actions*, and a *history* of observations and action execution is considered; we do not currently need these capabilities in our work. The ground literals in an *answer set* obtained by solving Π represent beliefs of an agent associated with Π. All reasoning (e.g., planning and inference) can be reduced to computing answer sets of Π [14]. We use the SPARC system [15] to compute answer set(s) for inference with observations.

ASP-based representation of knowledge has some advantages. It supports *default negation* (negation by failure) and *epistemic disjunction*. Unlike "$\neg$a", which implies that "*a is believed to be false*", "not a" only implies that "*a is not believed to be true*"; unlike "p $\vee \neg$p" in propositional logic, "p or $\neg$p" is not tautological. Also, unlike classical first-order logic, ASP supports non-monotonic logical reasoning, i.e., adding a statement can reduce the set of inferred consequences, aiding in elegant recovery from errors due to the incomplete knowledge. Modern ASP solvers support efficient reasoning in large knowledge bases, and are used by an international community.

D. Combined Model and Other Relations

The QSR-based and MSR-based groundings may disagree on the spatial relation between objects in any given scene. The control node then chooses the more reliable grounding to determine the spatial relation. This choice is based on a simplistic strategy that initially assigns high (low) confidence to the QSR-based (MSR-based) grounding and then updates the relative confidence based on the number of times the output from each grounding matches human feedback—we use integers (e.g., $\in [0, 10]$) that are incremented or decremented to represent the confidence levels. Object shapes and sizes may also influence spatial relations depending on the viewpoint. However, since the MSR-based grounding is based on histograms of relative distances and angles, it can be used to infer spatial relations over a range of viewpoints.

There are some caveats related to learning approach described above. First, notice that the QSR-based grounding is assumed to be reasonably accurate initially; if this assumption does not hold, an entirely inaccurate MSR-based grounding may be learned, resulting in incorrect estimates of spatial relations. Second, human feedback improves the specialized grounding (MSR) and overall accuracy of estimating spatial relations, but it is not essential for estimating spatial relations. Third, the encoded prepositions (with learned groundings) can be used to infer other spatial relations of interest. For instance, in our architecture, the spatial relation on may be defined by the following axiom:

$$on(Obj_1, Obj_2) \leftarrow above(Obj_1, Obj_2), \; touching(Obj_1, Obj_2). \tag{5}$$

which states that if object Obj_1 is above Obj_2 and touching it, then Obj_1 is on Obj_2. There has also been some recent work on learning such axioms interactively [16]. Finally, although we currently assume that each pair of objects can be related through one position-based and one distance-based spatial relation, not all the prepositions under consideration are mutually exclusive.

IV. EXPERIMENTAL RESULTS

In this section, we describe the experimental setup and the results of experimental evaluation.

A. Experimental Setup

For experimental evaluation, we used the benchmark Table Object Scene Database (TOSD)[3] and simulated scenes. This database contains 111 scenes for training and 131 scenes for testing—many scenes include complex object configurations, e.g., Figure 1a, while some scenes have only two objects, e.g., Figure 5a. Since TOSD includes segmentation labels but not spatial relation labels, we manually labeled 200 scenes for experiments. The simulated scenes were generated with a real-time physics engine (Bullet physics library). To generate these scenes, the ground truth definitions of different spatial relations were provided manually. Different subsets of 21 household objects from the Yale-CMU-Berkeley (YCB) dataset [17], along with a table and a shelf, were then used to create 1400 simulated scenes (200 for each preposition). An additional 25 labeled scenes for each preposition (175 total) were used for training. We tested two hypotheses:

- **H1** the proposed approach enables more effective use of human feedback;
- **H2** the combination of (manually-encoded) QSR grounding and the automatically-learned MSR grounding performs better than either grounding used individually.

As the performance measure, we used the accuracy of the labels assigned to spatial relations between pairs of objects. We also qualitatively evaluated the ability to support easy identification and correction of errors. In the description below, *all claims of improvement are statistically significant at* 95% *significance level*.

B. Experimental Results

The first set of experiments was designed as follows, with the results summarized in Table I:

1) Pairs of objects extracted from the training set of the TOSD were divided into 10 subsets.
2) Seven pairs of objects from each subset were used to train the MSR model with human feedback. Each pair represents one of the position-based spatial relations (*in, left, right, front, behind, above, below*).
3) Seven pairs of objects from each subset labeled with human feedback, and 200 pairs (excluding the seven with human feedback) with relations labeled using QSR grounding, were used to train the MSR model.
4) The MSR model trained using QSR and human feedback was used along with the QSR model, with the control node choosing between the two models.

[3]https://repo.acin.tuwien.ac.at/tmp/permanent/TOSD.zip

TABLE I: Comparison of (a) MSR grounding trained with just human feedback; (b) MSR grounding trained with 200 pairs labeled by the QSR grounding and seven pairs labeled with human feedback; and (c) the combination of MSR, trained as in (b), and QSR-based grounding with the choice made by the control node.

Training sets	Accuracy of labels over test set of 200 object pairs		
	MSR (feedback)	MSR (QSR + feedback)	Combined model
Sets 1	65%	77%	84%
Sets 2	82%	80%	94%
Sets 3	68%	80%	85%
Sets 4	66%	83%	87%
Sets 5	65%	74%	82%
Sets 6	68%	77%	86%
Sets 7	64%	87%	90%
Sets 8	64%	84%	91%
Sets 9	62%	82%	87%
Sets 10	52%	72%	81%
Mean	65%	79%	87%
Std Dev	7.2%	4.6%	8.3%

The three schemes ($\#2, \#3, \#4$ above) were evaluated on 200 object pairs in the test scenes of varying complexity. The results indicate that MSR models trained by QSR make better use of human feedback than MSR models trained using just human feedback, which supports hypothesis **H1**. Also, the control node-based combination of the MSR and QSR provides better accuracy than just MSR.

The second set of experiments was designed as follows, with the results summarized in Table II:

1) Pairs of objects extracted from the training set of the TOSD were divided into five subsets.
2) A MSR model was trained using QSR-based labels for four out of the five subsets ($\approx$ 2000 pairs) in each run.
3) The combination of the MSR model (trained as above) and the QSR model, with the choice made by the control node, was also considered.

The two different schemes ($\#2, \#3$ above) were evaluated on a set of 200 object pairs in scenes of varying complexity (ground truth, once again, was obtained manually). The results in Table II indicate that the combined model provides more accurate estimates of spatial relations than the models based on QSR or MSR (individually), which supports hypothesis **H2**. Recall that these claims were tested for statistical significance.

Table III shows experimental results for simulated scenes, using QSR-based labels as the baseline for comparison. We evaluated MSR-based groundings trained with varying amount of human feedback (i.e., no QSR)—we used one, 15, and 25 training sets, each with seven object pairs. The corresponding models were tested on 1400 object pairs from simulated scenes of varying complexity.

Next, we conducted experiments similar to those summarized in Table I but using larger number of simulated images for training and testing. The MSR model trained just

TABLE II: Comparison of (a) only QSR model; (b) MSR model trained by $\approx$ 2000 pairs labeled with QSR (but no human feedback); and (c) combination of MSR (as trained in (b)) and QSR with the choice made by the control node.

Training sets	Accuracy of labels over test set of 200 object pairs		
	QSR only	MSR trained by QSR	Combined model
Sets 1+2+3+4	70%	62%	96%
Sets 1+2+3+5	70%	62%	96%
Sets 1+2+4+5	70%	60%	95%
Sets 1+3+4+5	70%	60%	96%
Sets 2+3+4+5	70%	60%	96%
Mean	70%	61%	96%
Std Dev	0	1.1%	0.5%

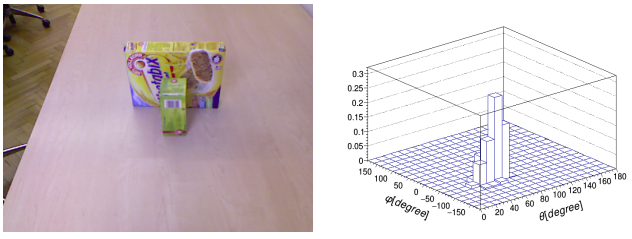
TABLE III: Comparison of QSR model with MSR models trained using only human feedback.

Model	Accuracy of labels over test set of 1400 object pairs
QSR	61.9%
MSR after 1 training set	96.1%
MSR after 15 training sets	98.5%
MSR after 25 training sets	98.6%

with human feedback had an accuracy of 95.9%, whereas the MSR model trained by QSR and human feedback had an accuracy of 97.2%. These results are similar to those obtained with the TOSD.

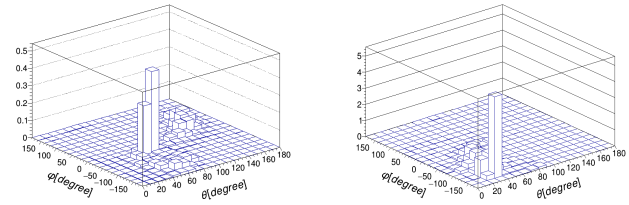
Most errors with our combined model correspond to truly ambiguous spatial relations between objects, e.g., a scene in which object A can be considered to be to the "left" or "behind" object B. Errors with other models are due to models being (or becoming) inaccurate—all results are order-independent. We then evaluated the ability to identify and correct errors with the learned MSR-based grounding. Figure 5a shows an TOSD image for which the MSR models incorrectly stated that the larger box is *above* the smaller one. We compared the learned visual words for the incorrect label and correct label ("behind") with the histogram extracted from the object pair in the image.

The χ^2 measure (Equation 2) between the learned and observed visual words was 0.325 for *above* and 0.319 for *behind*. In this case, even the QSR-based grounding detected



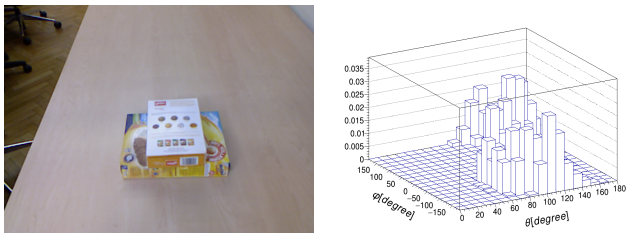
(a) TOSD scene. (b) 2D histogram.

Fig. 5: (a) Image from TOSD dataset; (b) Histogram generated from the image considering the smaller box as a reference.



(a) Histogram for *above*. (b) Histogram for *behind*.

Fig. 6: Histograms representing learned MSR groundings for: (a) above; and (b) behind.



(a) Scene for *above*. (b) Corresponding histogram.

Fig. 7: (a) Image with one object *above* another; and (b) corresponding histogram.

349 points in the *above* region and 23 in the *behind* region. In other words, the error seems to be because the QSR-based labels provided incorrect input to the MSR model.

Next, we compared the actual histogram between the objects in the test image, as shown in Figure 5b, with the histograms from the MSR grounding for *above* and *behind*—Figure 6. We observed that the visual word for the objects in the image is similar to the learned grounding for *above*. Under normal viewpoints and camera orientation, the θ angle is greater than 90° for *above*, but most points in the corresponding learned visual word correspond to $\theta \approx 60^\circ$.

To correct this error, we used an image that truly contains an instance of the *above* relation—Figure 7a. As expected, most of θ values in the revised visual word were $\in [90^\circ, 120^\circ]$—Figure 7b. The MSR model's grounding of *above* then provided the correct spatial relations between the objects in Figure 5a—the χ^2 similarity scores were 0.319 for *behind* and 0.088 for *above*. This example illustrates how the architecture supports the identification of, and incremental recovery from, errors caused by variations in factors such as viewpoint and orientation.

V. CONCLUSIONS

Robots assisting humans need to be able to recognize, reason about, and provide understandable descriptions of spatial relations between domain objects. We described an architecture that uses Answer Set Prolog to represent and reason with incomplete domain knowledge that includes spatial relations computed and inferred using generic grounding of the spatial relations (QSR). Spatial relations estimated using QSR or obtained from humans (when available), are used to incrementally learn a more specialized grounding of the spatial relations. In parallel, a relative measure of confidence in the two groundings is computed, which is then used to choose between the two groundings to estimate spatial relations between objects in previously unseen scenes. Experimental evaluation on a benchmark dataset of tabletop images and on simulated scenes of furniture indicate promising results even when a small number of images are used for training. Future work will consider datasets with more drastic changes in factors such as viewpoint, orientation and scale. We will also explore the learning of action models that include the learned spatial relations. Furthermore, we will investigate the use of this architecture on a physical robot assisting humans in complex indoor domains.

REFERENCES

[1] J. Ye, K. A Hua, Exploiting depth camera for 3D spatial relationship interpretation, in ACM Conference on Multimedia Systems, pages 151-161, 2013.

[2] K. Zampogiannis, Y. Yang, C. Ferm, and Y. Aloimonos, Learning the spatial semantics of manipulation actions through preposition grounding, in International Conference on Robotics and Automation, pages 1389-1396, May 2015.

[3] D. Elliot and A. P. De Vries, Describing images using inferred visual dependency representations, in Annual Meeting of the Association for Computational Linguistics, pages 42-52, 2015.

[4] S. Fichtl, D. Kraft, N. Krüger, and F. Guerin, Using relational histogram features and action labeled data to learn predictions for means-end actions, in IEEE/RSJ International Conference on Intelligent Robots and Systems (Workshop on Sensorimotor Contingencies for Robotics, 2015.

[5] O. Mees, N. Abdo, M. Mazuran, and W. Burgard, Metric learning for generalizing spatial relations to new objects, in IEEE/RSJ International Conference on Intelligent Robots and Systems, pages 3175-3182, Vancouver, Canada, September 24-28, 2017.

[6] A. Belz, A. Muscat, M. Aberton, and S. Benjelloun, Describing spatial relationships between objects in images in English and french, in Workshop on Vision and Language, pages 104-113, 2015.

[7] S. Guadarrama, L. Riano, D. Golland, D. Gouhring, Y. Jia, D. Klein, P. Abbeel, and T. Darrell, Grounding spatial relations for human-robot interaction, in International Conference on Intelligent Robots and Systems, pages 1640-1647, 2013.

[8] A. Thippur, C. Burbridge, L. Kunze, M. Alberti, J. Folkesson, P. Jensfelt, and N. Hawes, A comparison of qualitative and metric spatial relation models for scene understanding, in AAAI conference, pages 1632-1640, 2015.

[9] F. Ziaeetabar, E. E. Aksoy, F. Wörgöter, and M. Tamosiunaite, Semantic analysis of manipulation actions using spatial relations, in International Conference on Robotics and Automation, pages 4612-4619, 2017.

[10] R. Paul, J. Arkin, N. Roy, and T. Howard, Efficient grounding of abstract spatial concepts for natural language interaction with robot manipulators, in Robotics: Science and Systems (RSS), Ann Arbor, USA, June 18-22, 2016.

[11] A. Pronobis and R. Rao, Learning deep generative spatial models for mobile robots, in RSS Workshop on Spatial-Semantic Representation in Robotics, Cambridge, USA, July 16,2017.

[12] M. Shridhar and D. Hsu, Grounding spatio-semantic referring expressions for human-robot interaction, in RSS Workshop on Spatial-Semantic Representations in Robotics, Cambridge, USA, July 16, 2017.

[13] R. B. Rusu, Semantic 3D object maps for everyday manipulation in human living environment, KI - Kunstliche Intelligenz, 24(4):345-348, 2010.

[14] M. Gelfond and Y. Kahl, Knowledge representation, reasoning and the design of intelligent agents, Cambridge University Press, 2014.

[15] E. Balai, M. Gelfond, and Y. Zhang, Towards answer set programming with sorts, in International Conference on Logic Programming and Nonmonotonic Reasoning, Spain, September 15-19, 2013.

[16] M. Sridharan and B. Meadows, A combined architecture for discovering affordances, causal laws, and executability conditions, in International Conference on Advances in Cognitive Systems, USA, May 12-14, 2017.

[17] B. Calli, A. Wallsman, A. Singh, and S. S. Srinivasa, Benchmarking in manipulation research, IEEE Robotics and Automation Magazine, (September):36-52, 2015.